Project: Capstone Project 1: Milestone Report

Problem Statement:

Caterpillar (construction equipment manufacturer) relies on a variety of suppliers to manufacture tube assemblies for their equipment. These assemblies are required in their equipment to lift, load and transport heavy construction loads. We are provided with detailed tube specifications, components, and annual volume datasets. Our goal is to build and train a model that can predict how much a supplier will quote for a given tube assembly based on given supplier pricing, and use this information to further categorize assemblies and suppliers such that any movement in business criteria example recency, frequency, total spend, supplier rebates etc. can be accurately classified and responded with appropriate supplier strategy.

To solve this problem, project is divided in two subgroups 'Predict Supplier Pricing' and 'Categorize Assembly & Suppliers' to understand trends and clusters.

Description of Dataset:

Each supplier has their unique pricing model. Tube assemblies can vary across a number of dimensions, including base materials, number of bends, bend radius, bolt patterns, and end types. From buyer's perspective, altering any of these specs that are lower cost, requires fewer assembly steps and meets business requirements, will result in total lower cost.

Supplier price is quoted in 2 ways:

1. Bracket or Bulk Pricing: This is based on order quantity purchased
2. Non-Bracket or Non-Bulk Pricing: This is based on minimum order quantity.

This data set was obtained from public repository, Kaggle. Dataset contains 21 tables and 77% missing values which after treatment and feature engineering was reduced to 21,095 observations and 42 attributes. Our goal was to preserve relevant features and minimize information loss.

Data Cleaning steps included data conversion, fixing quantity mismatch pricing error, removing duplicates, reducing 2000+ categorical attributes, handling outliers and treating missing values. I used pandas merge and concatenation function to consolidate various tables into a single table for modelling.

I also added some new features such as total cost, area, rebates and segmentation features such as order supplier preference, market difficulty etc. to form clusters and drive insights from the raw data.

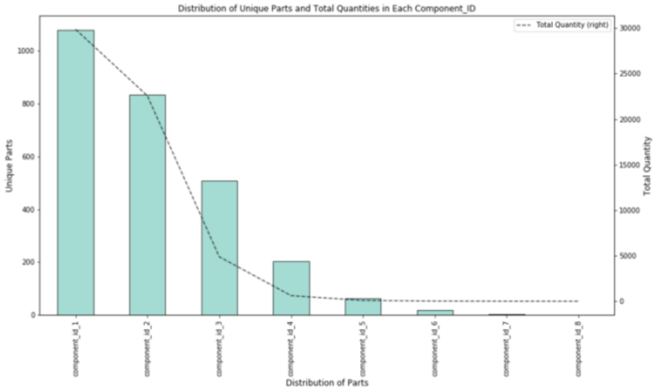
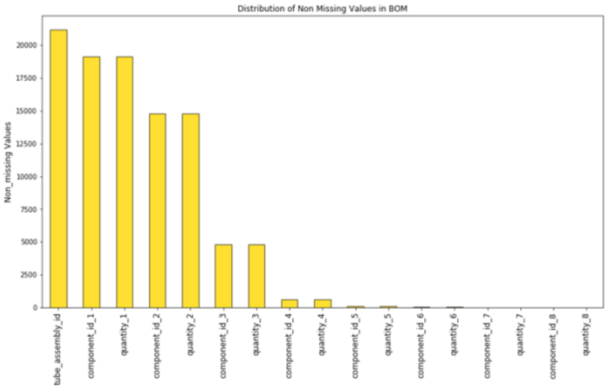
Summary of Findings: This is divided in to two sections.

Section 1: Supplier Pricing Prediction

Few Project Highlights related to supplier prediction:

1) Bill of Material Distribution & Missing Values:

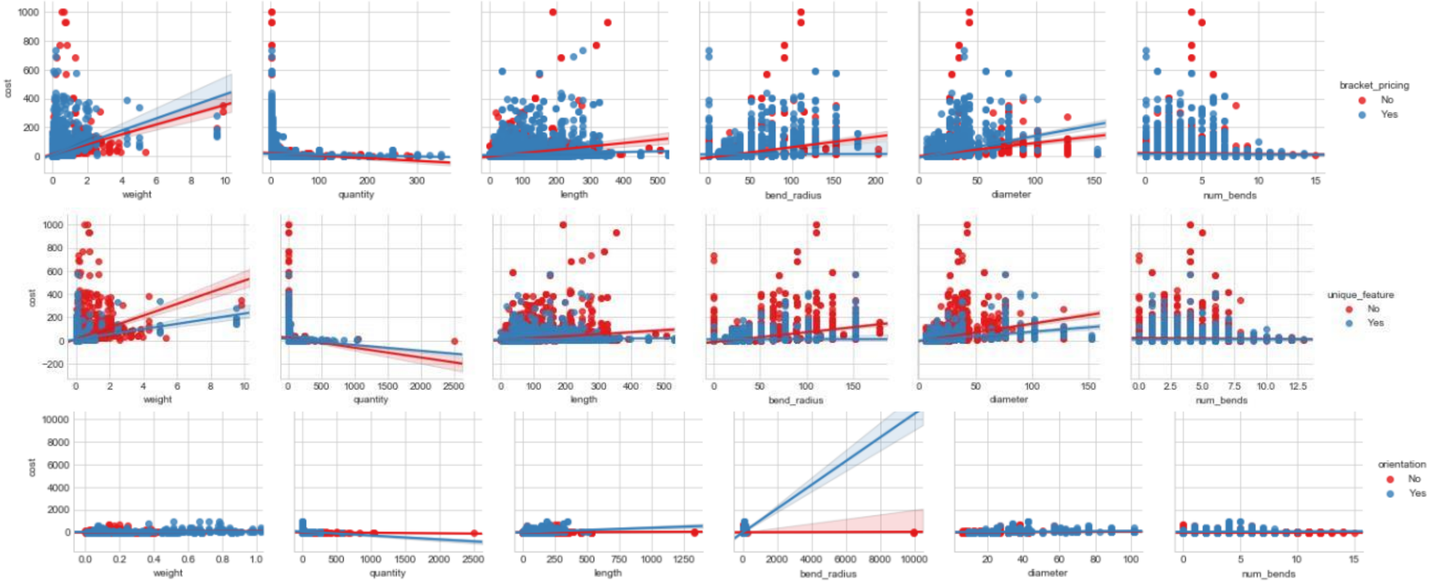
This shows what makes up the assembly, how many components, unique parts and quantities that would be required to manufacture these assemblies. By looking at the BOM distribution, it was suspected that some assemblies had missing components, which may need to be rejected as per below.

Graph 1: Distribution of Unique Parts & Quantities per Component	*Graph 2: Distribution of Non-Missing Values in BOM*
	

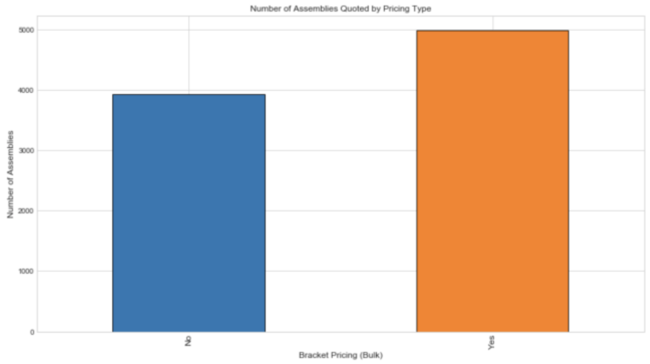
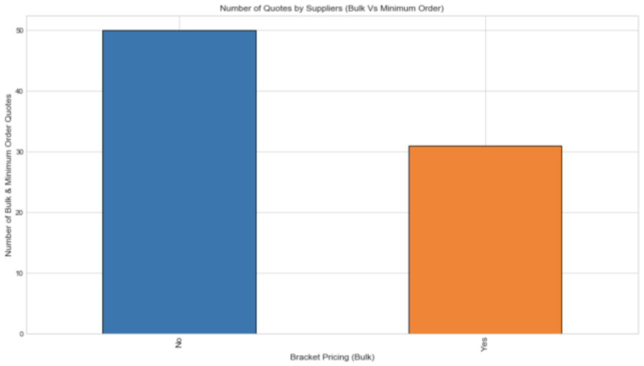
2) Supplier Cost Dependency:

Supplier price is dependent on tube specifications such as weight, length, diameter, bend radius etc. as well as pricing type such as bulk and minimum order quantity. It is also clear that suppliers offer exclusive pricing for both bulk and non-bulk assemblies as a result same assemblies will not have both bulk and non-bulk pricing. One main benefit of buying in bulk would be getting volume rebates.

a) Supplier Cost Dependency Graph:



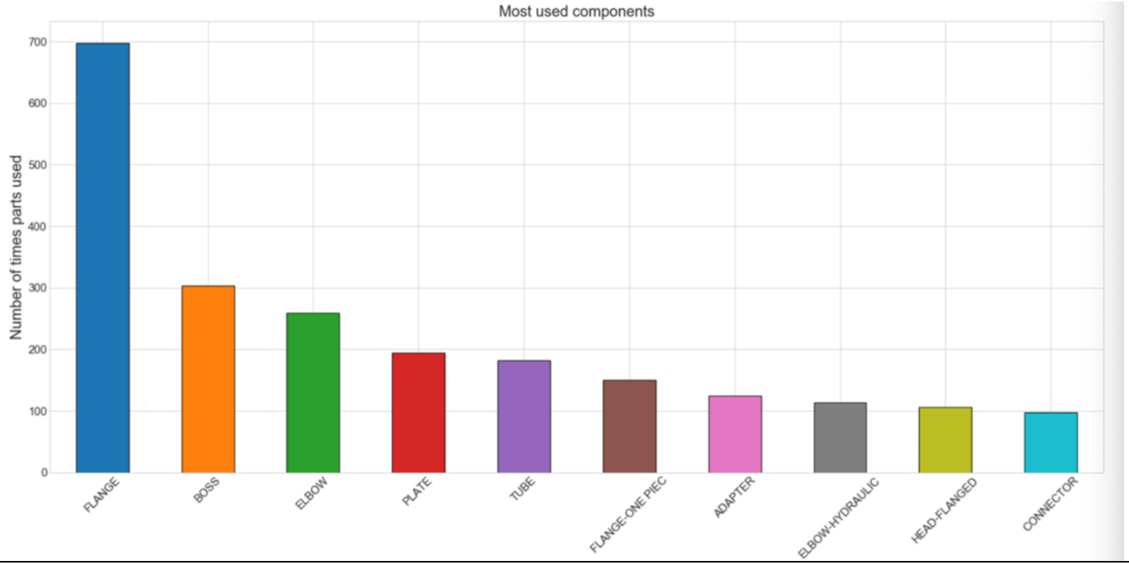
a) Supplier Pricing Distribution:

By Sole Suppliers (Either Bulk or Non-Bulk): Shows More Bulk Assembly Quotes Than Non_Bulk	By Multi-Suppliers (carrying both bulk and non-bulk): Shows More Non_bulk Quotes Than Bulk Quotes
	

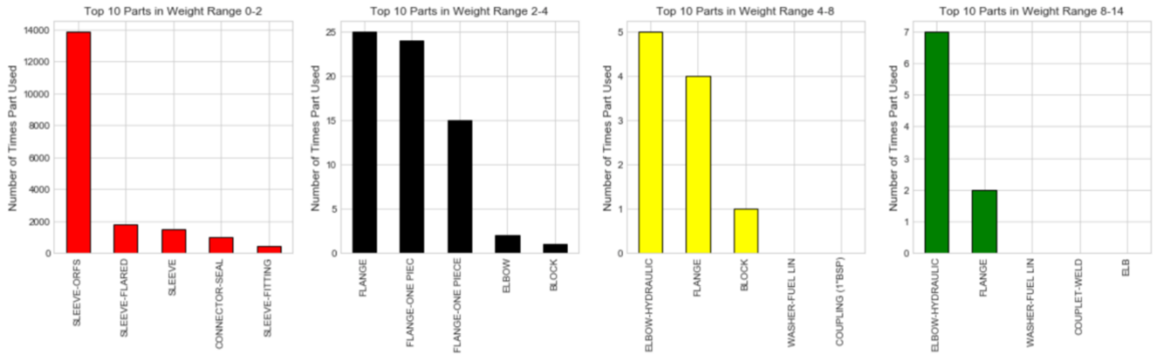
3) Most Used Components in Assembly Manufacturing:

From supplier cost dependency graph, it is observed that cost is dependent on the weight of the assembly. We identified overall distribution of most used components as well as by different weight ranges. This can be useful in comparing suppliers' total cost.

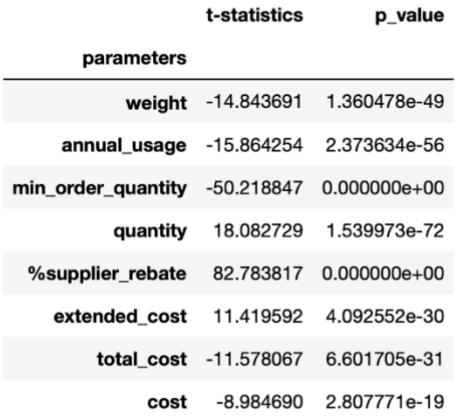
Overall Distribution of Most Used Components:



Most Used Components by Different Weight Range:

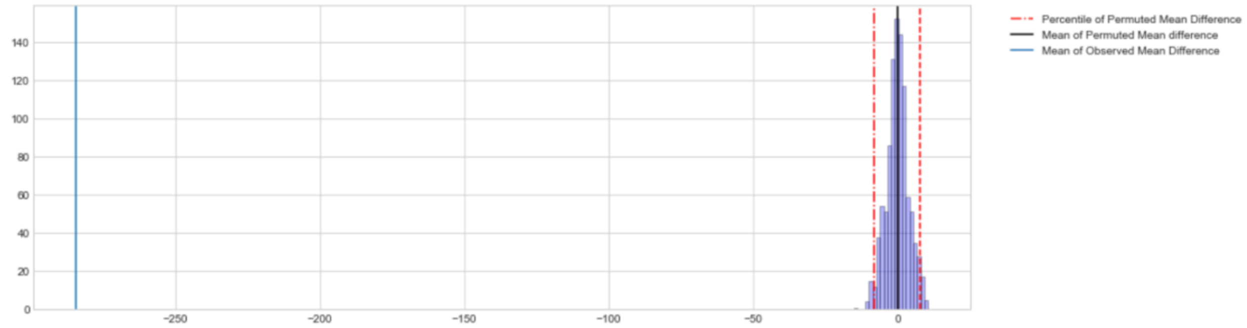


4) Difference Between Bulk and Non-Bulk Assemblies:
Our objective here is to understand the difference in bulk vs non_bulk assemblies using statistical measure. We used two-sided t-test to calculate test statistics and verify if p-value is < 0.05. Null was rejected and statistically seen that bulk and non-bulk assemblies comes in multiple different parameters, which are not common.



	t-statistics	p_value
parameters		
weight	-14.843691	1.360478e-49
annual_usage	-15.864254	2.373634e-56
min_order_quantity	-50.218847	0.000000e+00
quantity	18.082729	1.539973e-72
%supplier_rebate	82.783817	0.000000e+00
extended_cost	11.419592	4.092552e-30
total_cost	-11.578067	6.601705e-31
cost	-8.984690	2.807771e-19

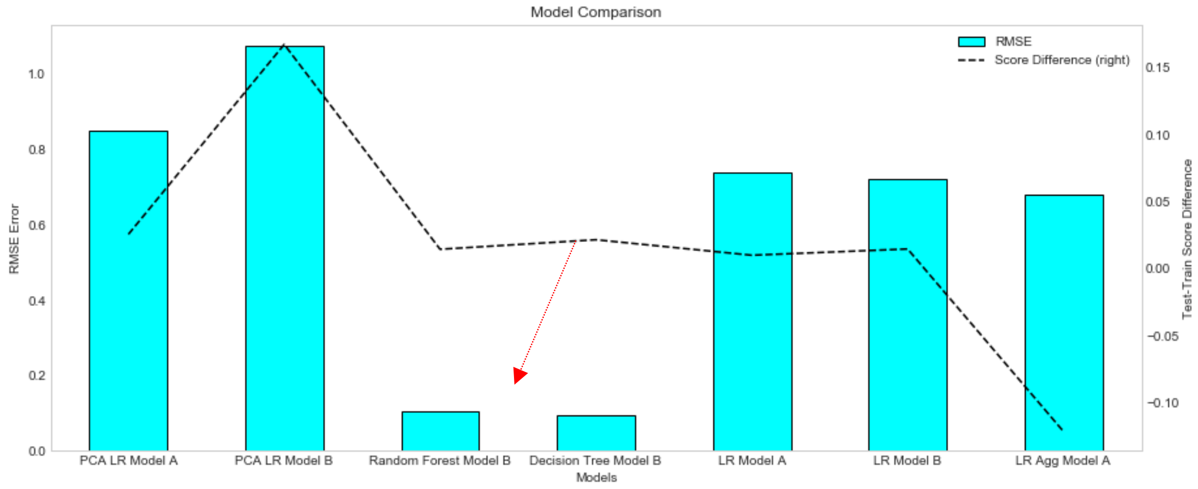
Additionally, if we collected more data, say 1000 samples, could there be an instance that business would like to buy bulk assemblies on a minimum order basis, and if yes, how much change in p-value would there be?



P-Value: 1.0, which is > 0.05
Confidence Interval of Mean of Permuted Mean Difference at 95% level: [-8.10, 7.93]
Mean of Observed Mean Difference: -284.16
Since, p-value > 0.05, null exists. Business will continue to buy bulk assemblies in bulk, unless mean of observed mean difference i.e. -284 reaches above the upper sample mean range of 7.93, which would be quite significant.

5) Model Prediction: To get started, I used few different models to compare test-train scores and rmse error as shown in graph below. I noticed that ensemble algorithm

resulted in best accuracy 98% and lowest rmse error on training and validation test dataset.



Section 2: Categorize Assembly and Suppliers

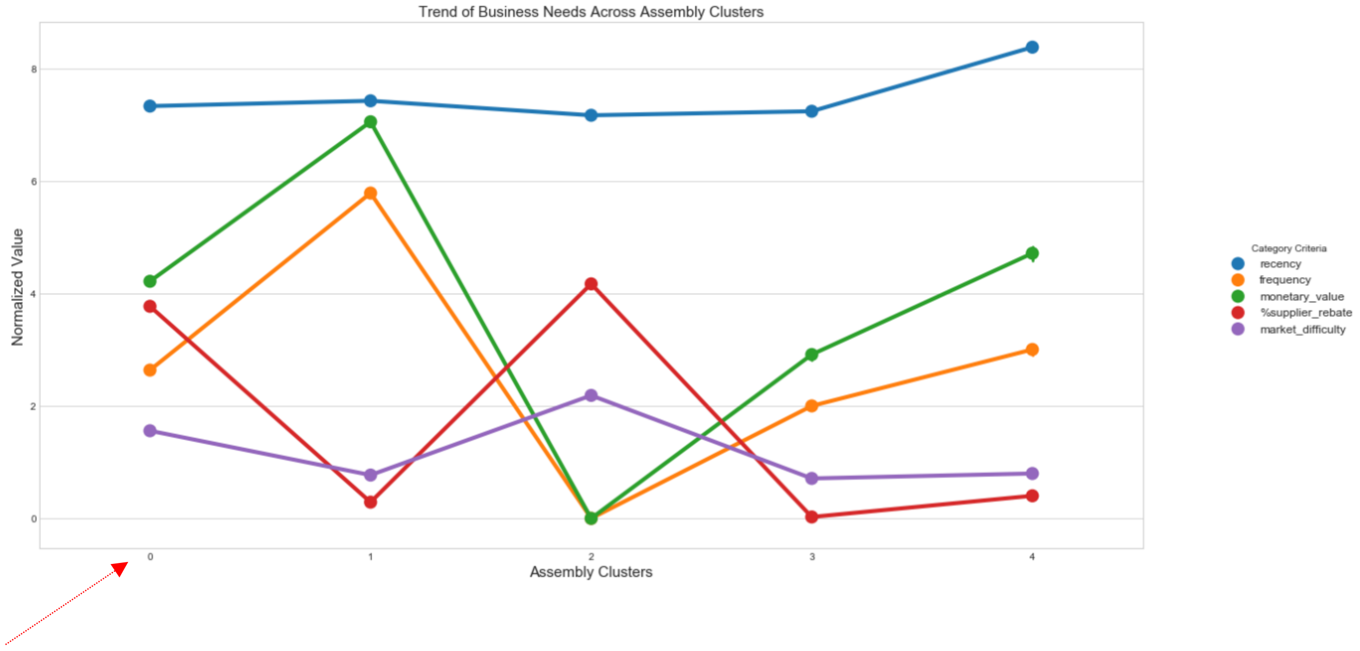
Few Project highlights related to clustering:

1) Pre-analysis: It was important to establish a pre-categorization baseline such as contract length, monetary spend, supplier preference etc. to understand and assess assembly and supplier trends.

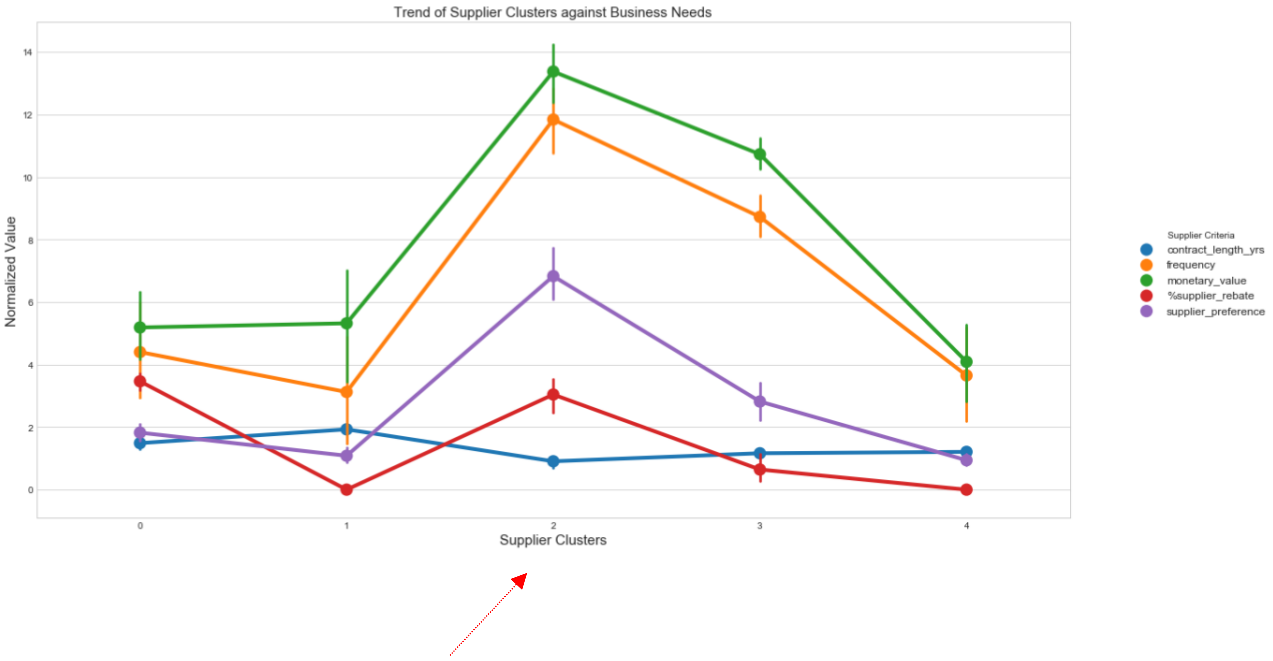
2) Algorithm: Then, applied clustering algorithm to re-group suppliers and assemblies using above criteria.

3) Analysis: Observed groups of assembly and supplier clusters showing trend against pre-established category criteria. This was useful in identifying benchmarked clusters, which can be used in learning and improving efficiency of the other clusters.

   a) **Best Managed Assembly Category: Cluster 0-** High Spend Most Rebated Assemblies



b) Best Managed Suppliers: Supplier Cluster 2- Most preferred suppliers handling high spend, high frequency orders and best rebates



Future Possibilities:

1) Add additional features such as on time delivery, safety performance, contract compliance and inventory to understand other supplier performance factors.

2) Develop ETL pipeline to see and optimize supply chain performance in real-time.

3) Use time series modelling to predict future supplier pricing.

------End------